MONDAY, JULY 24

projects, sqlite, openrefine, large and ugly data sets

PART ONE

PROJECTS.

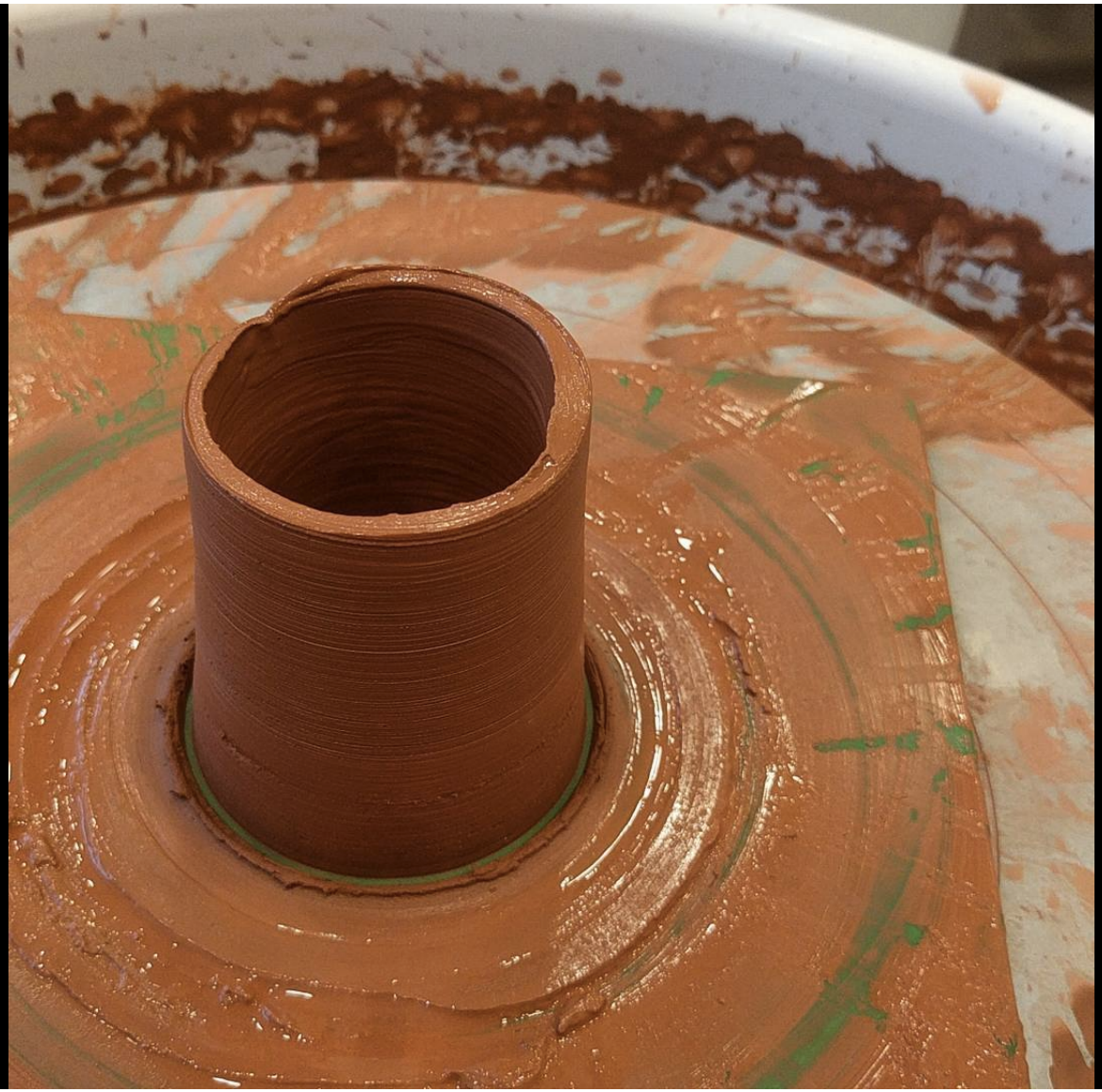
WHY ARE THESE IMPORTANT?

Findings are worthless without communication

Wasting everyone's time

You can't get better without failure, but you have to learn from those failures

It's seriously the only important thing we do



STAND UP

YOUR PROBLEM WAS...

WITH THE DATA?

WINDOW SIDE

WITH THE STORY?

DOOR SIDE

YOUR PROBLEM WAS...

FINDING STORY OR DATA?

CHALKBOARD

WORKING WITH DATA? VISUALIZING YOUR STORY?

BACK OF ROOM

1 OR 2 STORIES PER GROUP

YOU'LL PROVIDE YOUR ISSUE NUMBER AND WE'LL DISCUSS AS A CLASS

YOU HAVE 5 MINUTES TO FIND YOUR REPS

BUI'S CHEAP TRICKS

Central tendencies - mean, median, mode

Best and worst

Distribution

YOUR NEW MOTTOS

A bird in the hand is worth two in the bush

The perfect is the enemy of the good

Always be shipping

JFDI – Just Fucking Do It

NEW RULES

NO INTERACTIVES.

NO MAPS.

NOW LET'S HAVE CLASS.

Download the link from **#algorithms**!!!

"TOO BIG FOR PANDAS"
"...BUT I'M TOO LAZY FOR POSTGRES"

 SQLITE! 

"TOO MUCH DIRTY TEXT"
"TOO LAZY FOR PANDAS"

 OPENREFINE! 

INSTALLING SQLITE

OS X

```
brew install sqlite
```

Windows

Install chocolatey, the brew/apt for Windows, from **chocolatey.org**. Open a new prompt and

```
choco install sqlite
```

CAMPAIGN FINANCE DATA

1. Download
   ftp://ftp.fec.gov/FEC/2016/indiv16.zip
2. Visit
   http://classic.fec.gov/finance/disclosure/ftpdet.shtml

Why campaign finance data?

CAMPAIGN FINANCE DATA

1. Unzip both
2. Use **%%time**, and open 2016 data in pandas with **read_csv.** Be sure to give it **sep="|"** so it will open correctly.
3. How long did it take? Can we speed it up? Reading in a subset?

"TOO BIG FOR PANDAS"
"...BUT I'M TOO LAZY FOR POSTGRES"

 SQLITE! 

"TOO BIG FOR PANDAS"
"...BUT I'M TOO LAZY FOR POSTGRES"

 SQLITE3! 

LOADING DATA INTO SQLITE

first, cd into the same directory as your data

READING CSV INTO SQLITE

```
.mode csv
.separator ','
.import indiv_header_file.csv contributions
.mode csv
.separator '|'
.import itcont.txt contributions
```

READING SQL INTO PANDAS

```
import sqlite3
import pandas as pd
conn = sqlite3.connect("contributions.db")
df = pd.read_sql("select * from contributions
where STATE = 'NY'", conn)
df.head()
```

INDEXING?